Exploration of 2015 Yellow Taxicab Dataset

Brief description of the data set and a summary of its attributes

TLC Trip Record Data has 12 years (2009 –2020) worth of Data available but I have decided to analyze a subset of the 2015.

In 2015, passengers took nearly 300 million yellow cab rides in New York City. Working with the complete dataset for all these rides would require considerable time and computational resources. The 12 data files used represent two percent of the total trips sampled at random from each month.

I chose this data because it is useful for real-world applications such as:

- Predicting taxi duration for a trip
- Allocating taxi to zones/regions based on demand.

Below is the summary of the data attributes as described here

Attribute / Column	**Description**
VendorID	A code indicating the TPEP provider that provided the record. **1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.**
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged.
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged.
RateCodeID	The final rate code in effect at the end of the trip. **1= Standard rate** **2=JFK** **3=Newark** **4=Nassau or Westchester** **5=Negotiated fare**

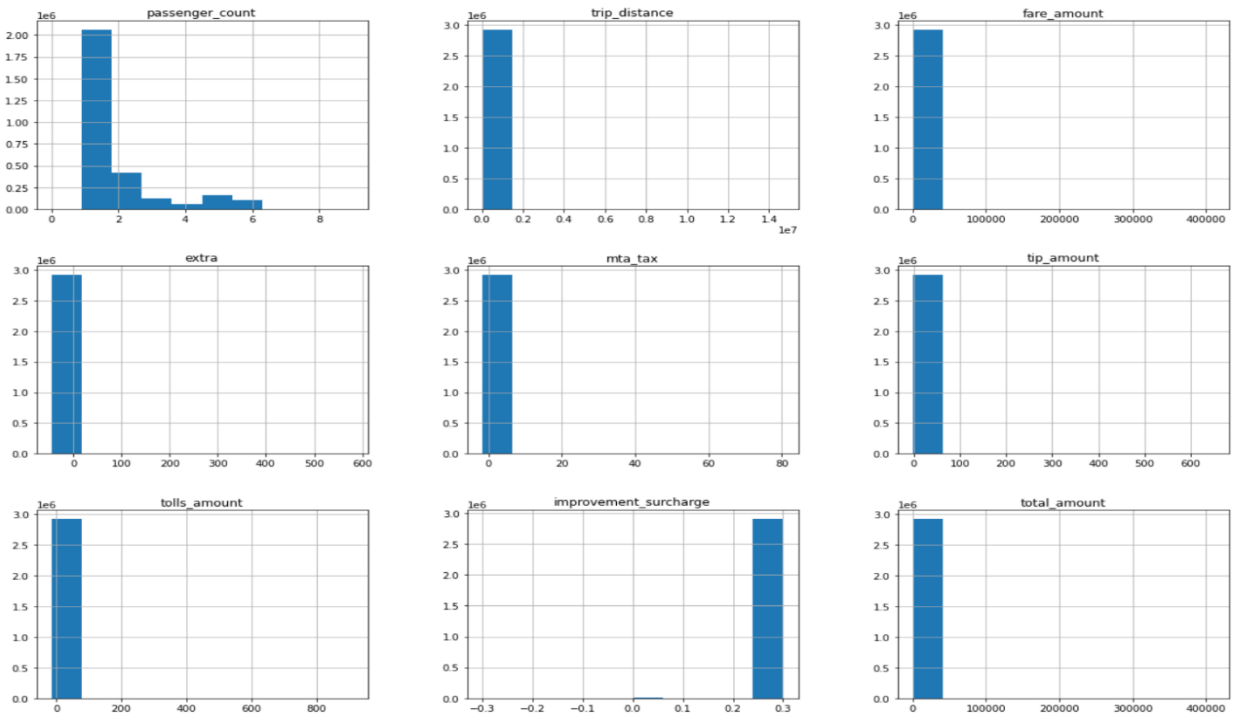
	6=Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. **Y= store and forward trip** **N= not a store and forward trip**
Payment_type	A numeric code signifying how the passenger paid for the trip. **1= Credit card** **2= Cash** **3= No charge** **4= Dispute** **5= Unknown** **6= Voided trip**
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
Tip_amount	Tip amount –This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

Data Exploration

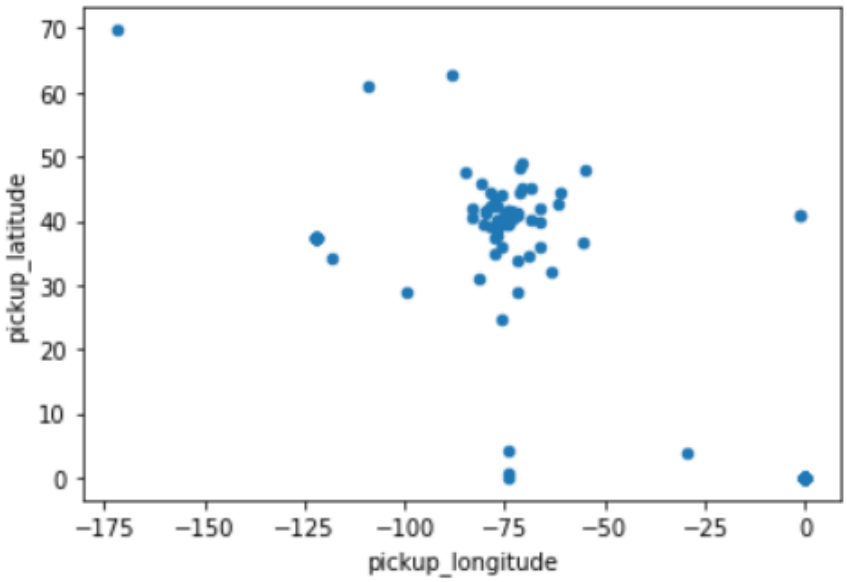
Joining the 12 files provided resulted in 2, 922, 266 instances in the dataset with no missing values. We have 12 numerical features, 4 categorical features, 2 datetime features and 1 integer. For easy visualization, I converted *VendorID, RateCodeID and payment_type* to their corresponding values.

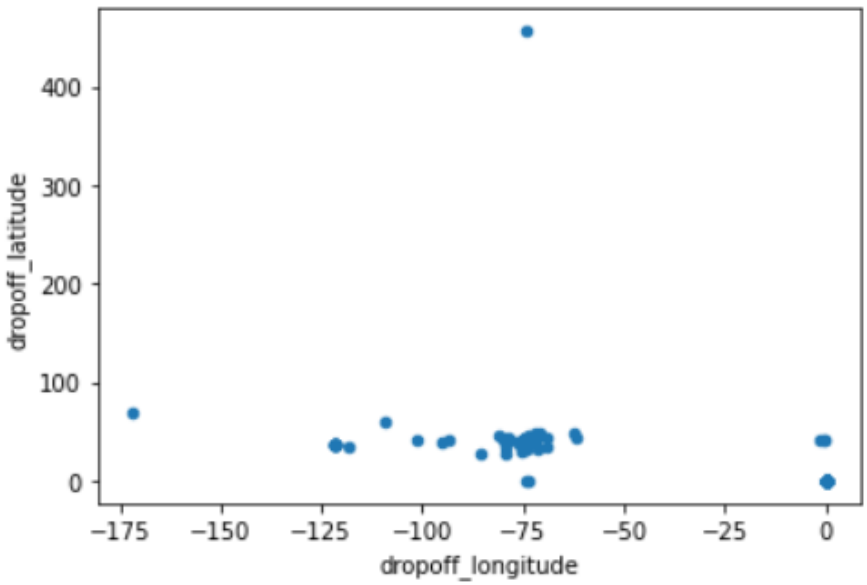
Summary of some the numerical attributes shows that this data is not perfect and therefore needs some cleaning for example there is no way we can have a negative tip or zero passengers.

	Passenger_count	**Trip_distance**	**Fare_amount**	**Tip_amount**
min	0.0	0.0	-150.00	-2.7
max	9.0	14680110.0	410266.86	650



Visualization of the geographical location also revealed some outliers.





In addition, RateCodeID contains a value of 99 which is clearly an error:

RateCodeID	Frequency
Standard rate	2845340
JFK	61564
Negotiated fare	9110
Newark	5046
Nassau or Westchester	1051
99	127
Group ride	28

Data Cleaning and Feature engineering

For data preprocessing/cleaning

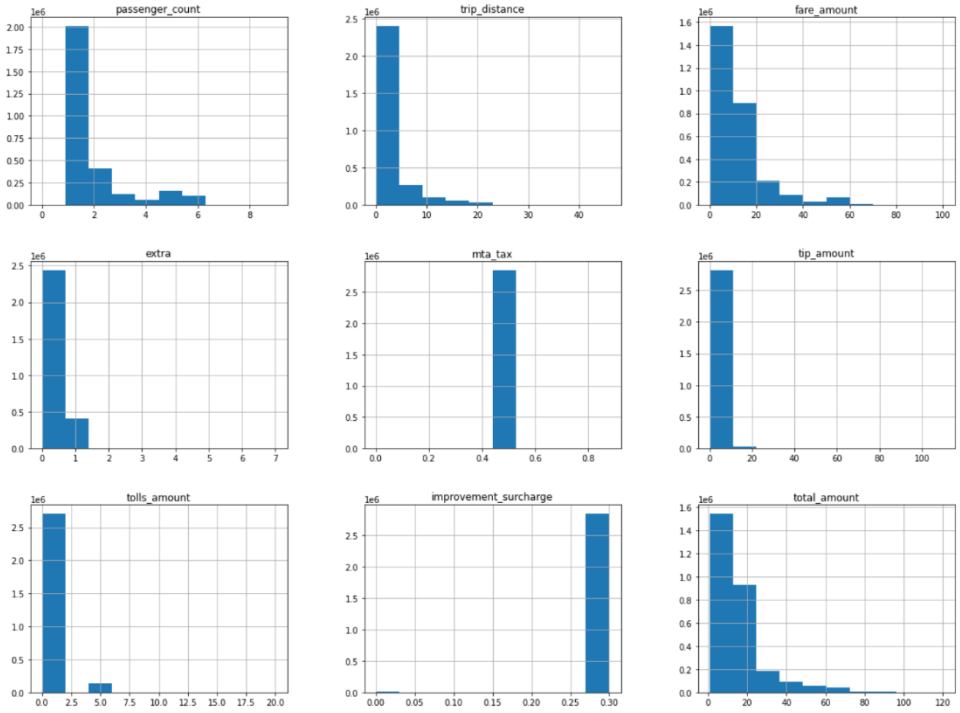
- Charges and trip information that are negative, as well as charges inconsistent with expected values are considered incorrect. In addition, trips with pickup or drop off locations outside a geographic region of interest are removed.
- Only keep trips with valid passenger and distance information.
- Remove trips missing valid pickup or drop off locations.
- Remove trips with invalid Rate code.

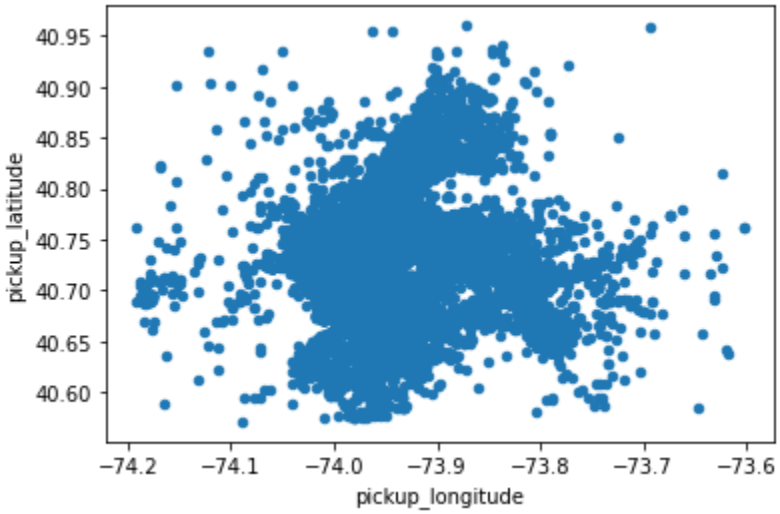
For Feature engineering, the following features were added.

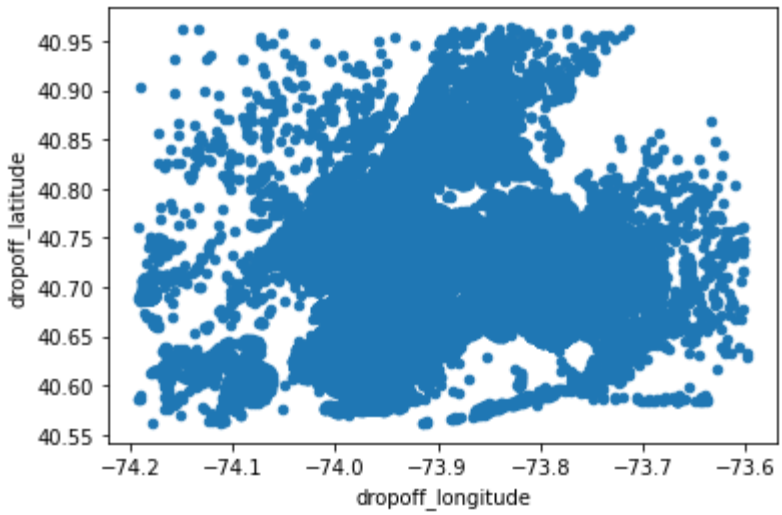
- ***duration*** - Length of the trip, in minutes, calculated from the pickup and drop off times.
- ***ave_speed*** - Average speed, in mph, calculated from the distance and duration values.

- ***time_of_day*** **-** This feature represents pick up time as the elapsed time since midnight in decimal hours (e.g. 7:10 am becomes 7.1667). The output is a duration vector with units of hours.
- ***day_of_week*** **-** This feature is a categoical array indicating the day of the week the trip began, in long format (e.g. 'Monday').

After removing invalid trip information, we can now visualize the features again







Key Findings and Insights

The table below shows how much each attribute correlates with total_amount : fare_amount, trip_distance, tip_amount, tolls_amount has strong positive correlation while mta_tax, dropoff_latitude, pickup_latitude has strong negative correlation with total_amount.

Features	Correlation with total_amount
total_amount	1.000000
fare_amount	0.982680
trip_distance	0.942129
tip_amount	0.696002
tolls_amount	0.692427
pickup_longitude	0.483289
dropoff_longitude	0.317986
passenger_count	0.012331
improvement_surcharge	0.006751
extra	-0.036715
dropoff_latitude	-0.199711
mta_tax	-0.241827
pickup_latitude	-0.270703

Next Steps in analyzing data

Use the Region and Zones dataset available to identify the zone and regions for each pickup and drop-off locations.

Get rid of attributes (e.g passenger_count, extra) that adds nothing to predicting total_amount.

Summary

This data set has very rich and important features that can be used to build a useful machine learning model for predicting total amount for a trip. In addition, after the initial cleaning steps we have about 2, 854, 781 rows of data which is reasonable amount of instance for any machine learning project.